

Network Science

dataScience UDD



A large, semi-transparent network graph is centered in the background. It consists of numerous small, glowing blue dots representing nodes, connected by a web of thin white lines representing edges. The graph has a organic, branching structure, with a denser cluster of nodes at the bottom left and more scattered connections towards the top right.

Cristian Candia-Castro Vallejos, Ph.D.

cristiancandia@udd.cl

Director Magister en Data Science UDD

Profesor Investigador, Facultad de Ingeniería, UDD

External Faculty Northwestern Institute on Complex Systems,
Kellogg School of Management, Northwestern University

Estructuras de Redes

Estas diapositivas se basan en la presentación original del Prof. Albert-László Barabási, de Northeastern University, con autorización.
El contenido ha sido traducido para su uso en este curso.

Conceptualización de una red

- **Dominio:** el área específica de estudio o aplicación para la cual la red será analizada.
- El dominio captura el conjunto de *nodos* y *enlaces* que forman una red en particular, así como también las características y propiedades de esa red.

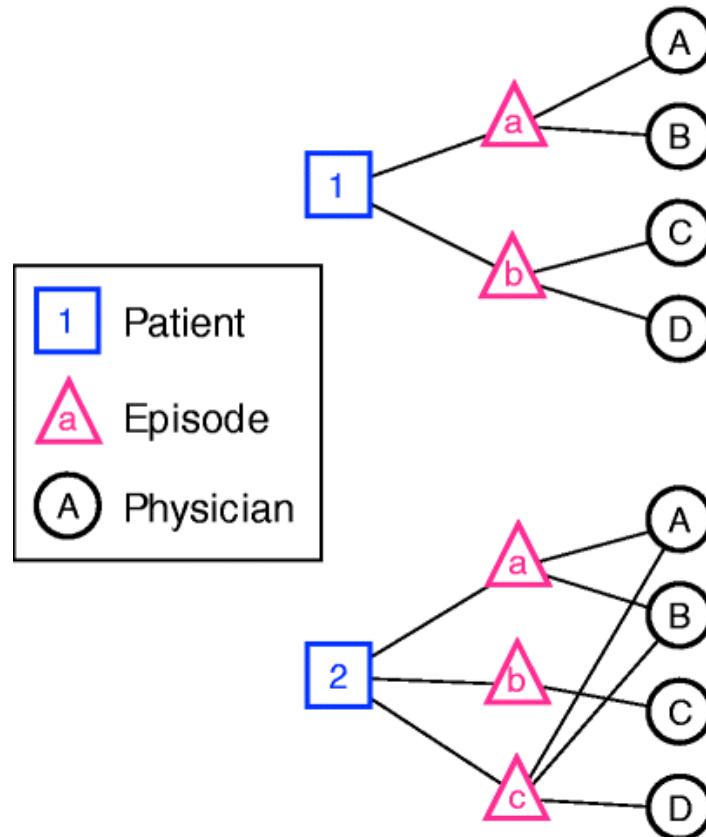
Conceptualización de una red

Preguntas a considerar:

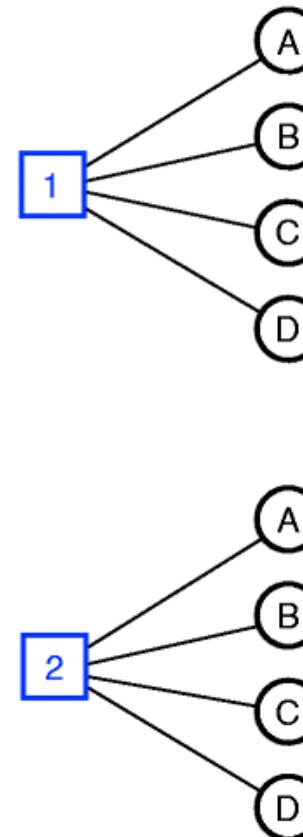
- **Qué sistema representa esta red?** E.g. sistema biológico, sistema computacional, sistema eléctrico, sistema social (personas).
- **Cuál es el significado de las relaciones de los nodos?** E.g. patrones de comunicación en una comunidad en particular, interacciones biológicas entre genes o proteínas.
- **Qué medidas de red son útiles para capturar las relaciones entre los nodos?** E.g. in-degree centrality para comunicaciones dirigidas entre individuos, betweenness o eigenvector centrality como indicador de prestigio social, detección de comunidades para capturar grupos de ideología política, flujo y vulnerabilidad en redes de electricidad.

Niveles de las Redes

Tripartite network



Bipartite network
(patient-physician)



GRAFOS BIPARTITOS

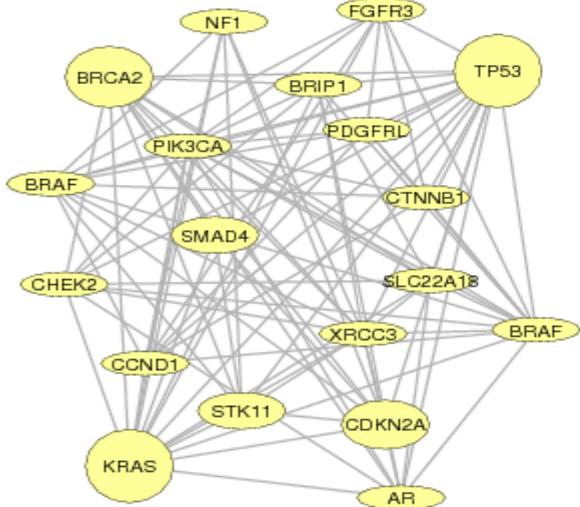
Un **grafo bipartito** (o bigrafo) es un gráfo cuyos nodos se pueden dividir en dos **conjuntos separados** U y V, de manera que cada enlace conecta un nodo en U con uno en V; es decir, U y V son conjuntos **independientes**.

Ejemplos:

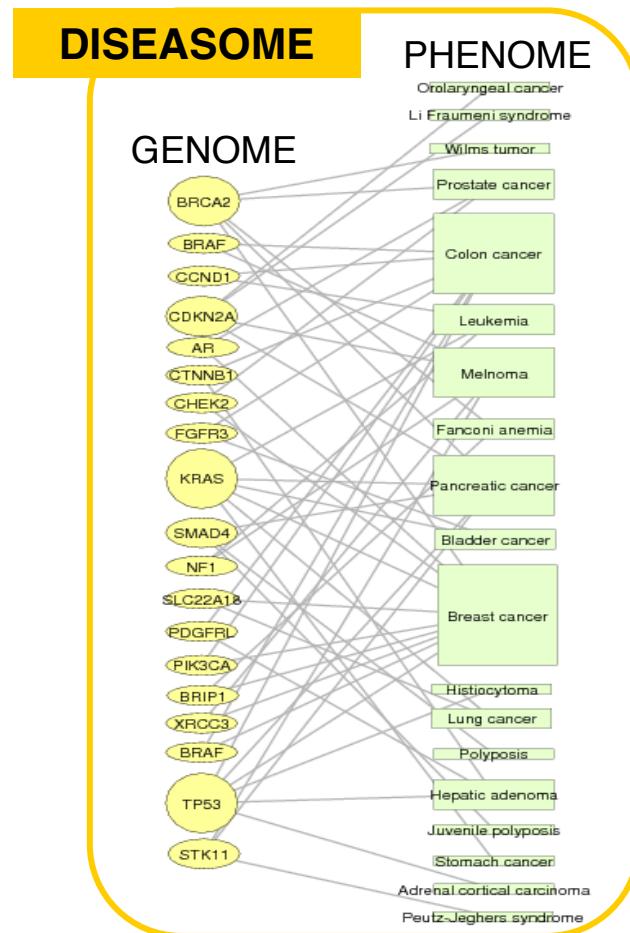
Red de actores de Hollywood Redes de colaboración

Red de enfermedades (diseasome)

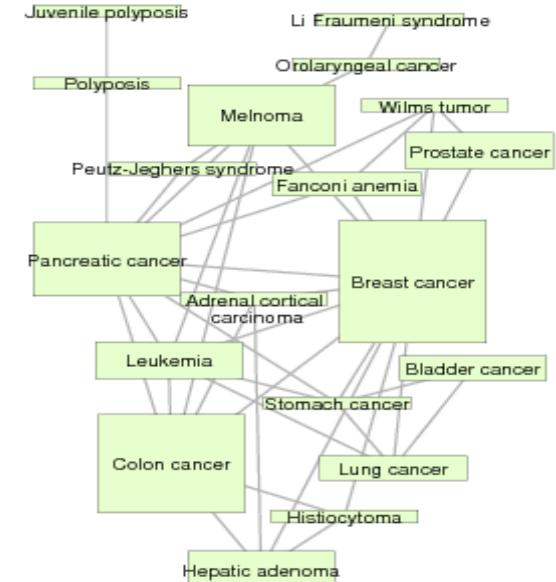
Red de genes– Red de enfermedades



Gene network



Goh, Cusick, Valle, Childs, Vidal & Barabási, PNAS (2007)

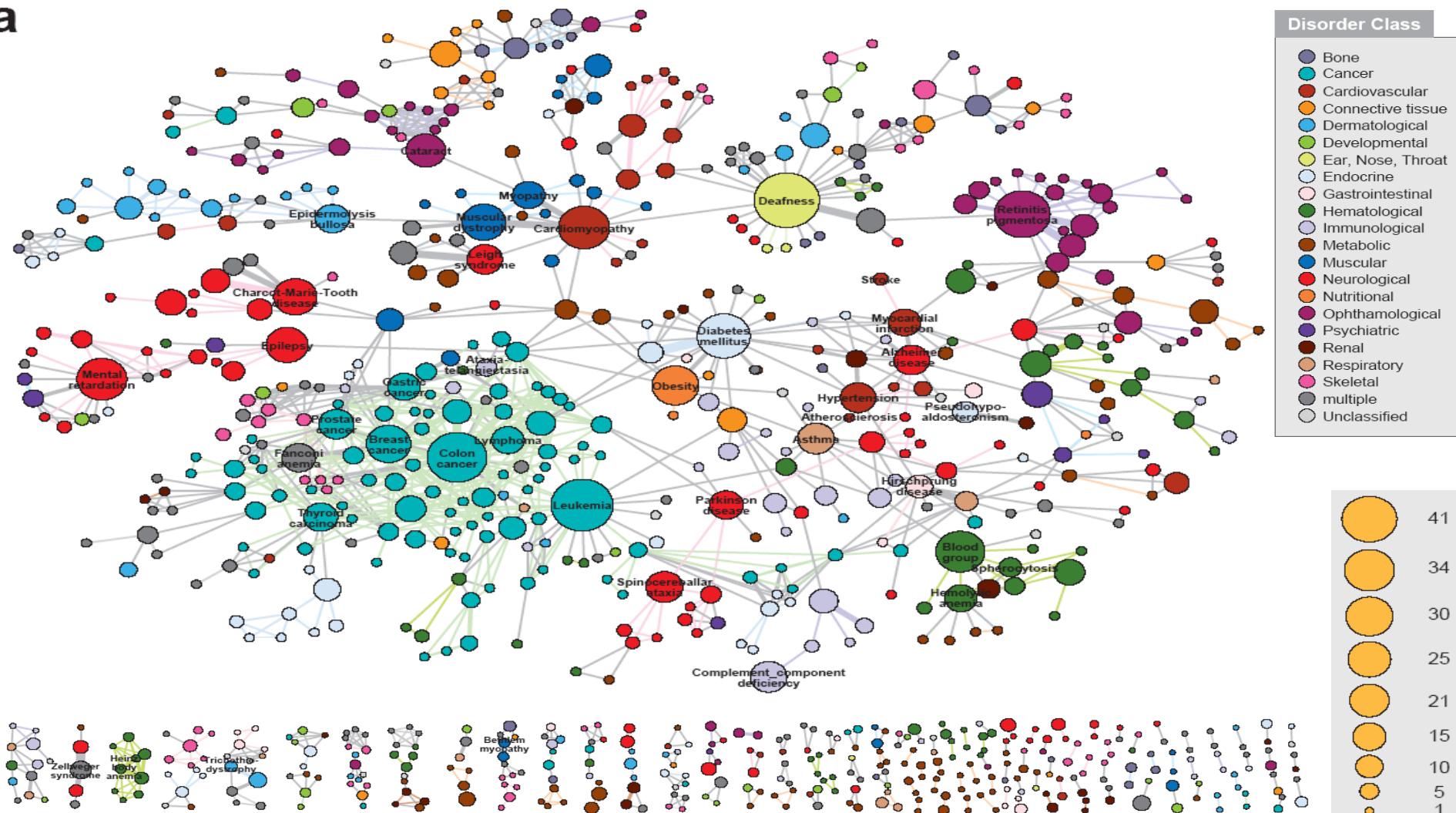


Disease network

RED DE ENFERMEDADES HUMANAS

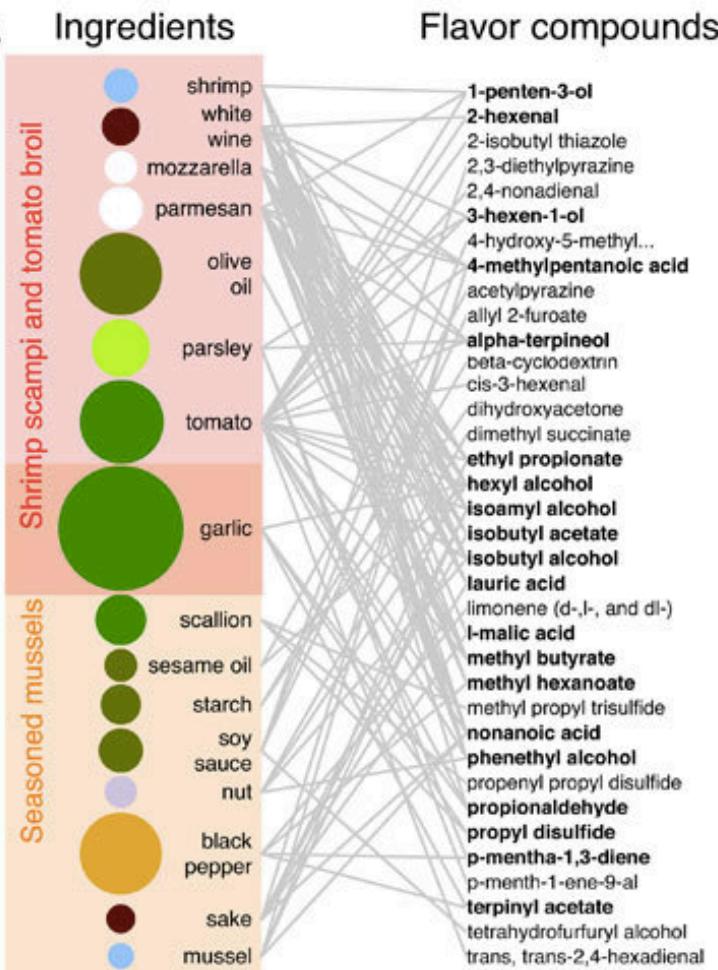
<https://www.nytimes.com/2008/05/06/health/research/06dise.html>

a

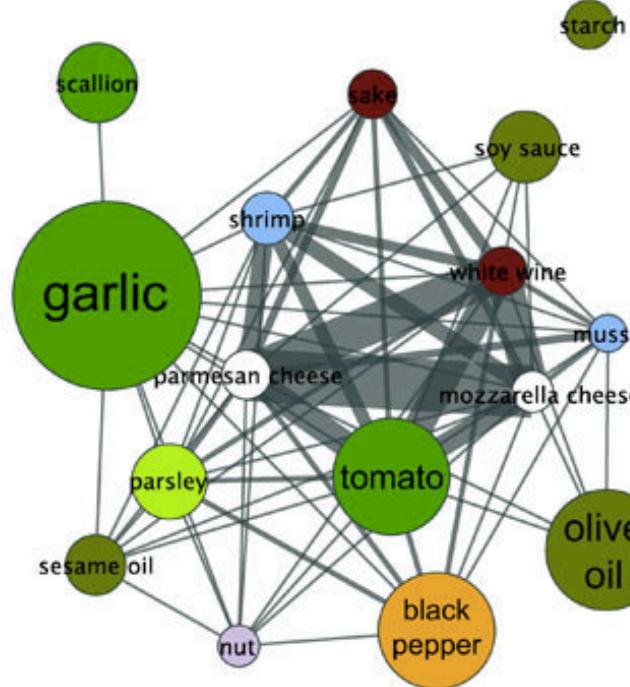


Red bipartita ingredient-sabor

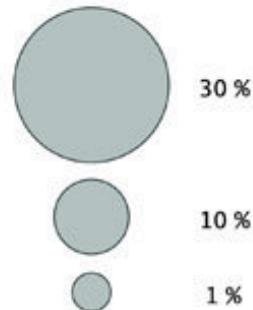
A



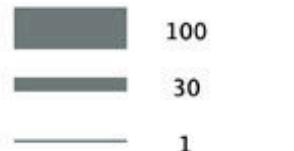
B Flavor network

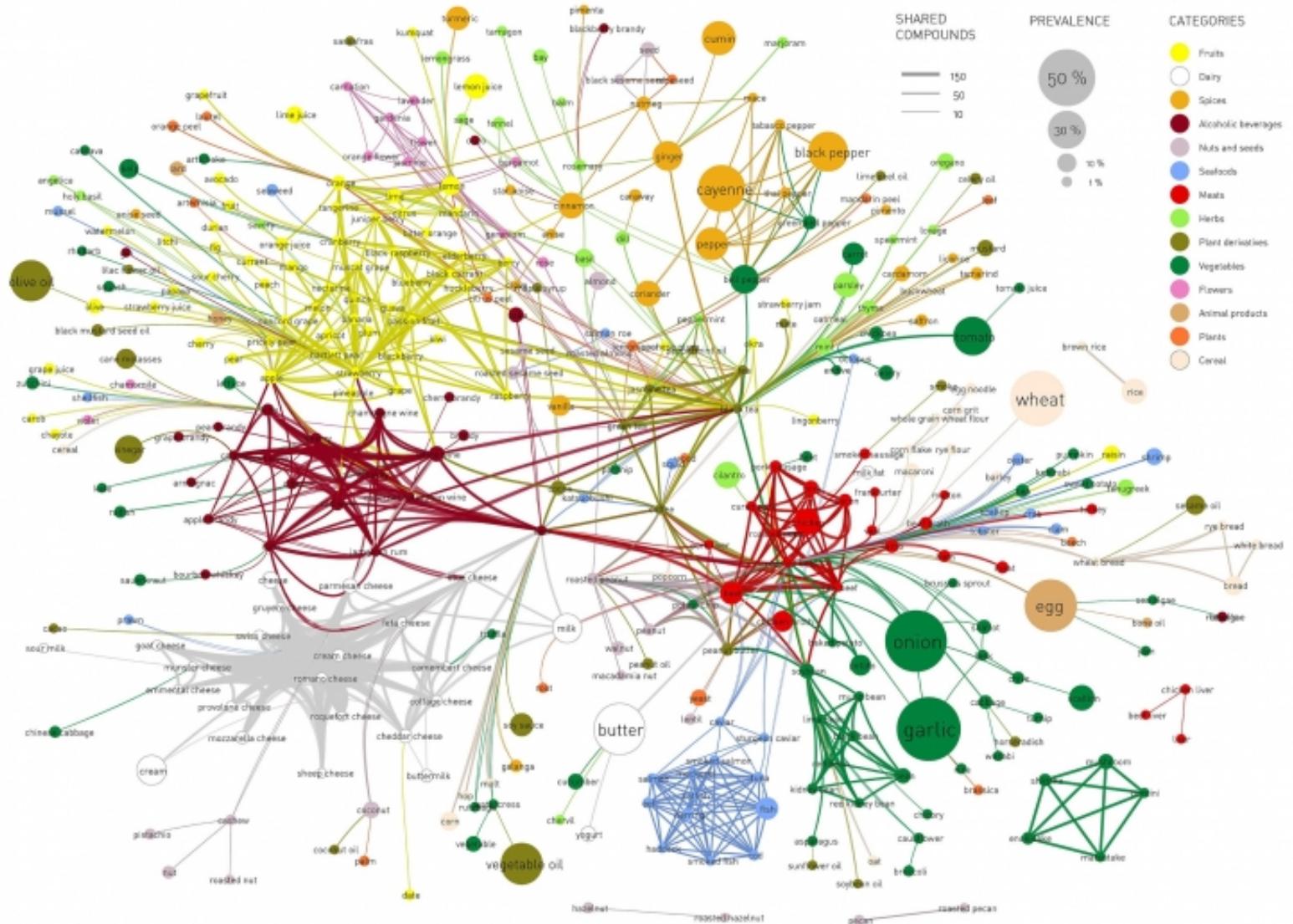


Prevalence



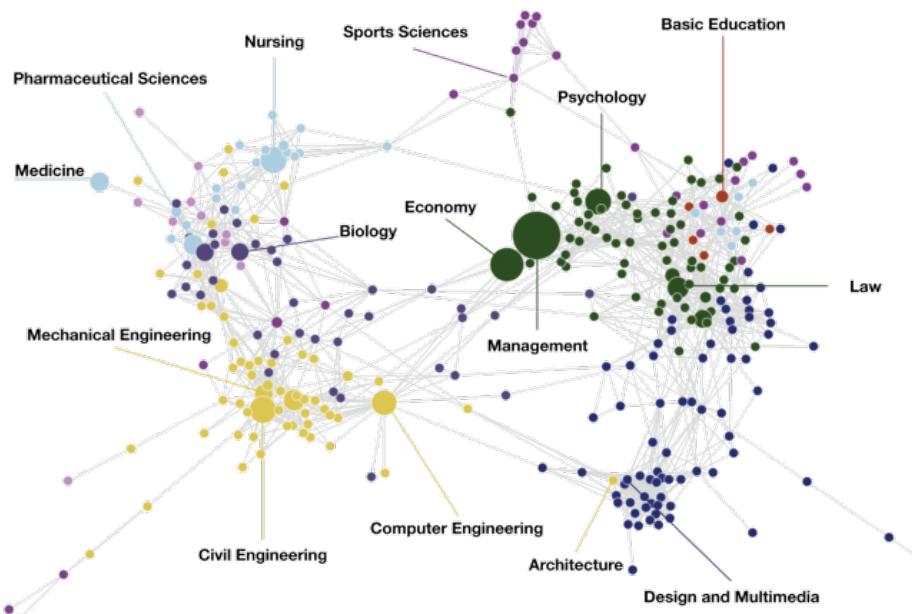
Shared compounds



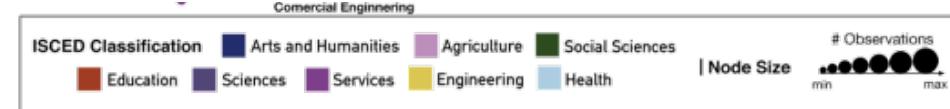
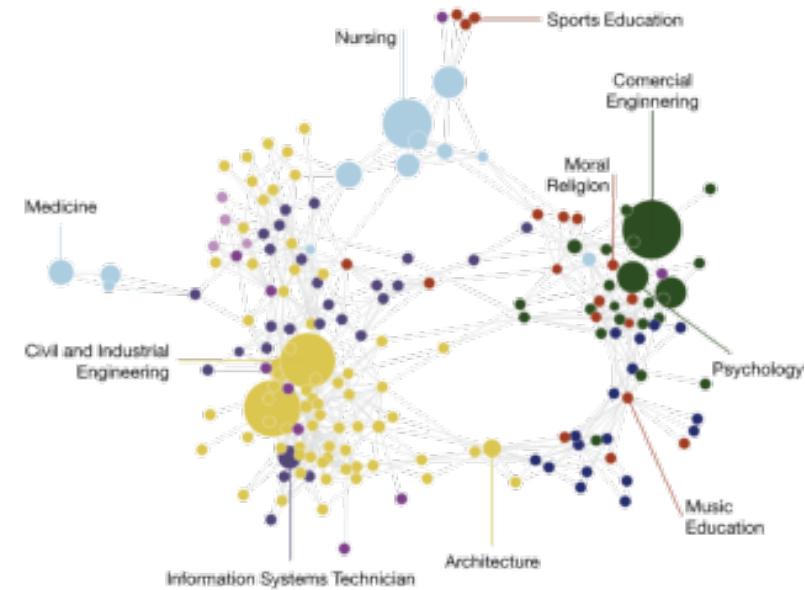
b.

Red bipartita carreras-postulantes

a) Portuguese Higher Education System [2008-2015]

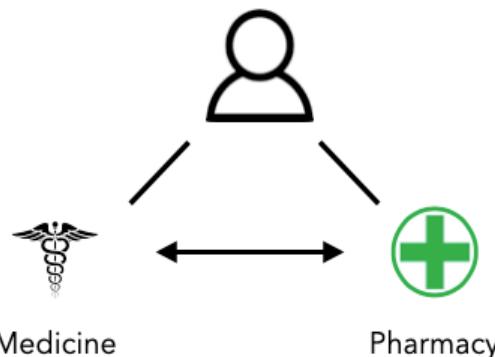


c) Chilean Higher Education System [2012-2017]



The Higher Education Space

Connecting Degree Programs



Lista de Preferencias

1. Medicina
2. Odontología
3. Tecnología Médica
4. Odontología
5. Cs. Físicas y Astronómicas



Pares de Carreras

Medicina	●—●	Odontología	●—●	Odontología
Medicina	●—●	Tecnología Médica	●—●	Cs. Físicas y Astronómicas
Medicina	●—●	Odonotología	●—●	Odontología
Medicina	●—●	Cs. Físicas y Astronómicas	●—●	Cs. Físicas y Astronómicas
Odontología	●—●	Tecnología Médica	●—●	Cs. Físicas y Astronómicas

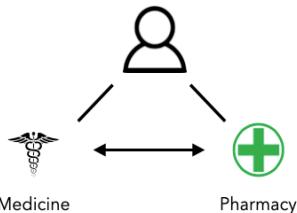
Consideramos todas las preferencias de cada postulante.
Las preferencias repetidas indican una postulación a dos
instituciones de educación superior distintas.

Creamos todos los pares de carreras posibles. Luego, descartamos los que contienen la misma carrera en ambos extremos (color gris).

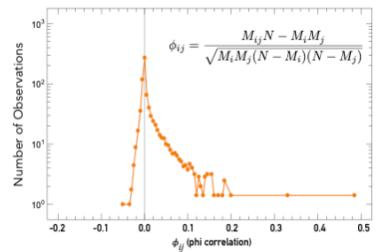
Red bipartita carreras-postulantes

The Higher Education Space

Connecting Degree Programs

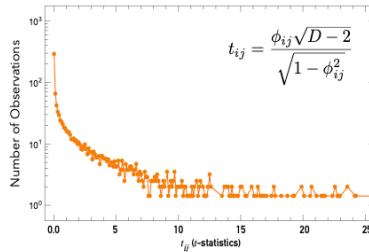


Finding Correlations Between Degrees

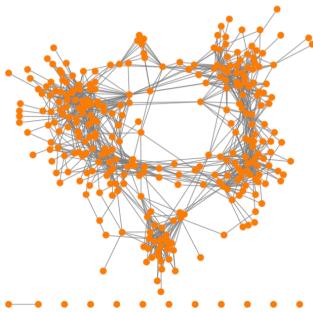


Discard all negative correlations

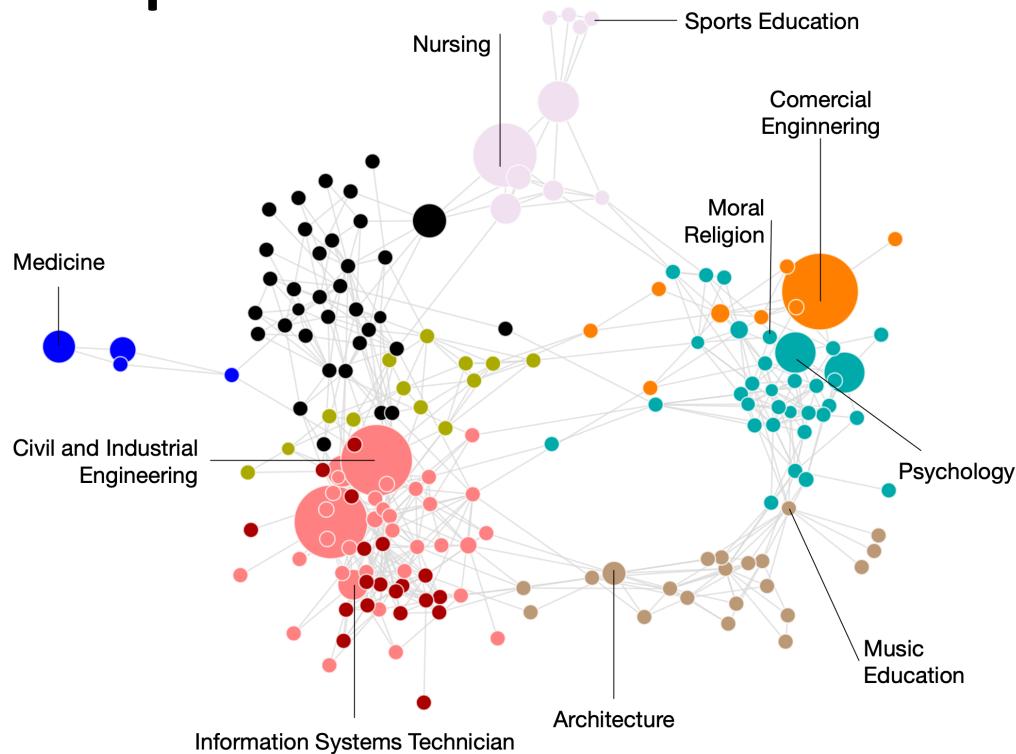
Finding Statistical Significance of Correlations



Discard all non-significant links

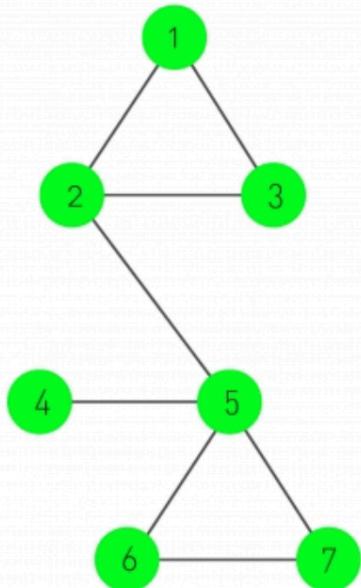


Discard all loose Nodes

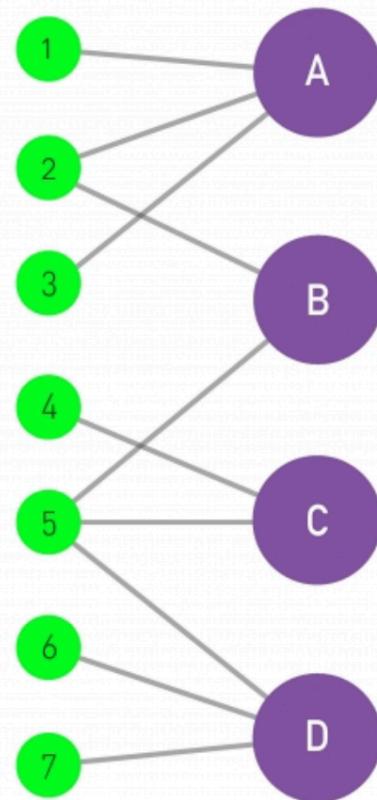


Proyecciones de una red bipartita

PROJECTION U U



U V



PROJECTION V

Ejemplo

Each node represent an *author* $i = \{1, 2, 3, \dots, n\}$ where n is the number of *authors* in the network G . The network of authors is represented as $G = (I, V)$, where I is the set of authors (nodes, k) and V is the set of connections between authors (edges, v).

There is an edge v between two authors $i_1, i_2 \in I$ so that $v(i_1, i_2) \in V$ if they collaborated in the same scientific publication.

Authors' productivity is measured using their *degree centrality*.

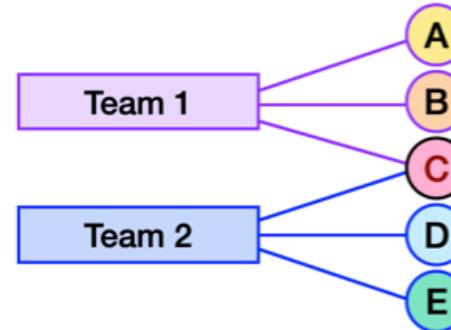
a

Original data

	Team 1	Team 2
Author 1	A	D
Author 2	B	E
Author 3	C	C

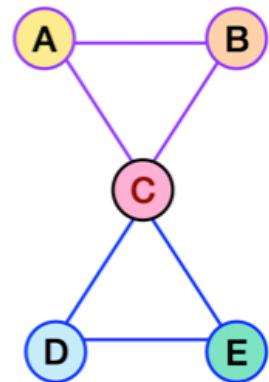
b

bipartite



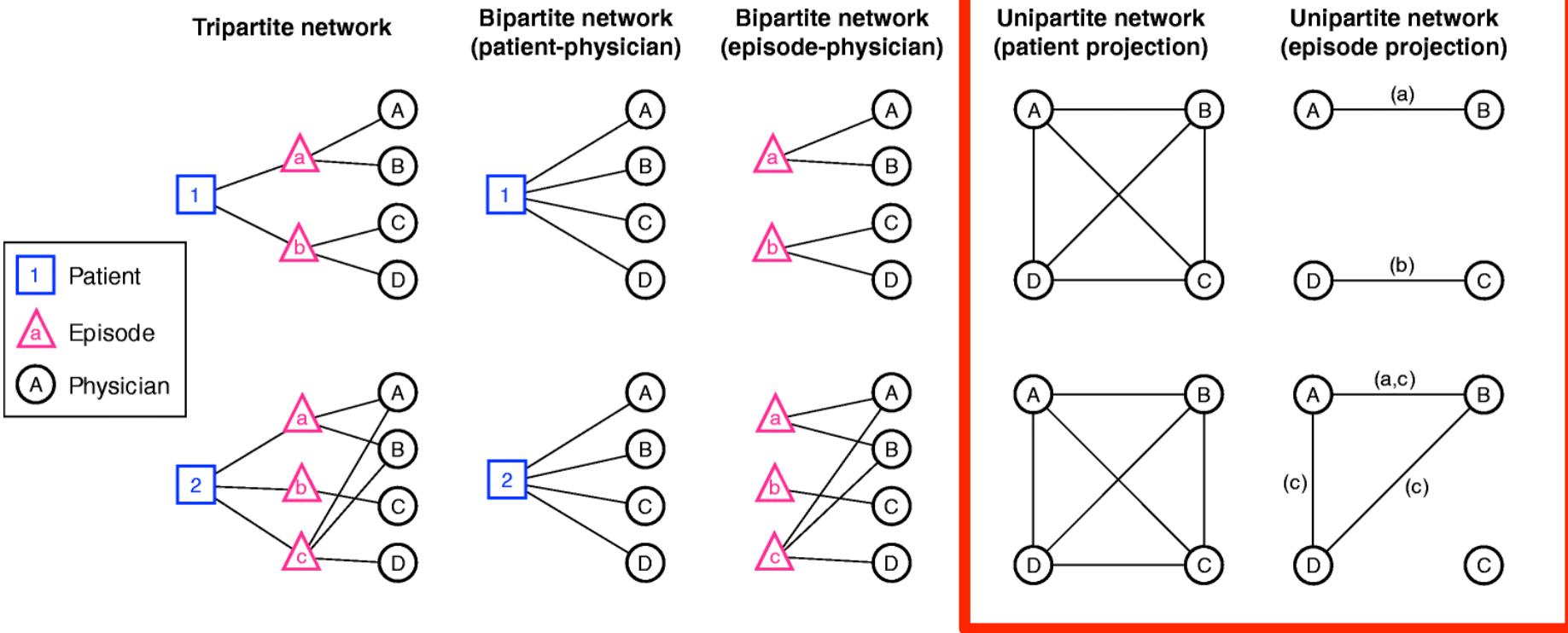
c

unipartite
(projection)



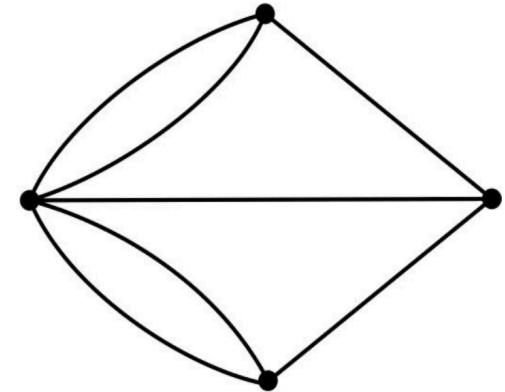
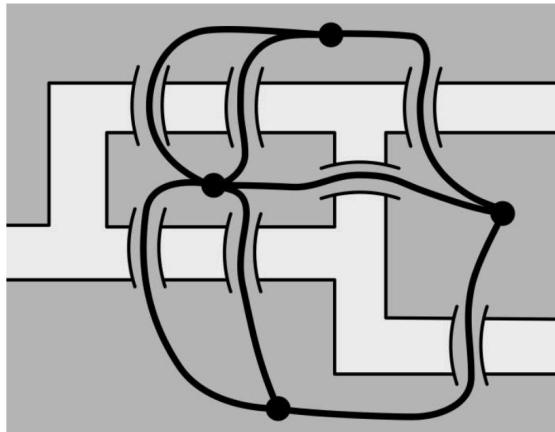
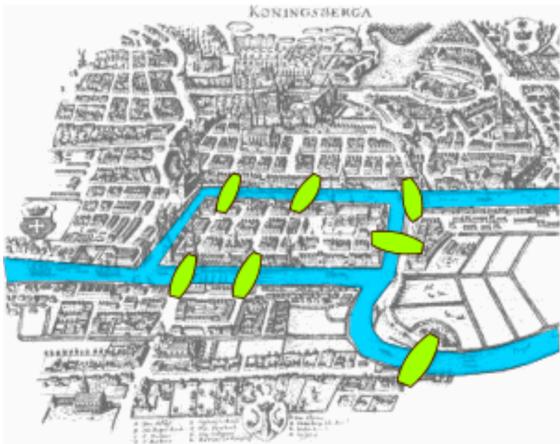
In this example, connections are observable

Proyecciones de diferentes niveles



Visualización de Estructuras de Red

Convertir una situación del mundo real, a un modelo abstracto con nodos y enlaces

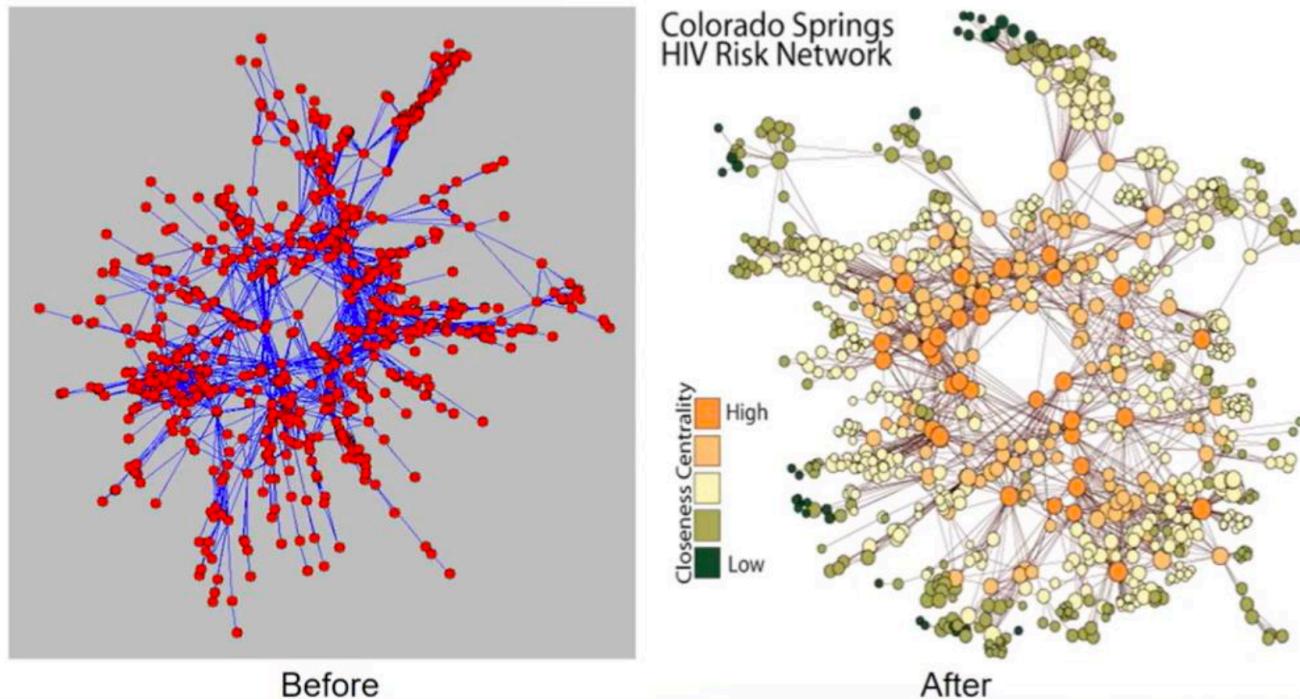


Source: <https://divisbyzero.com/2008/09/25/konigsberg-today/>

Elementos de visualización de la red:

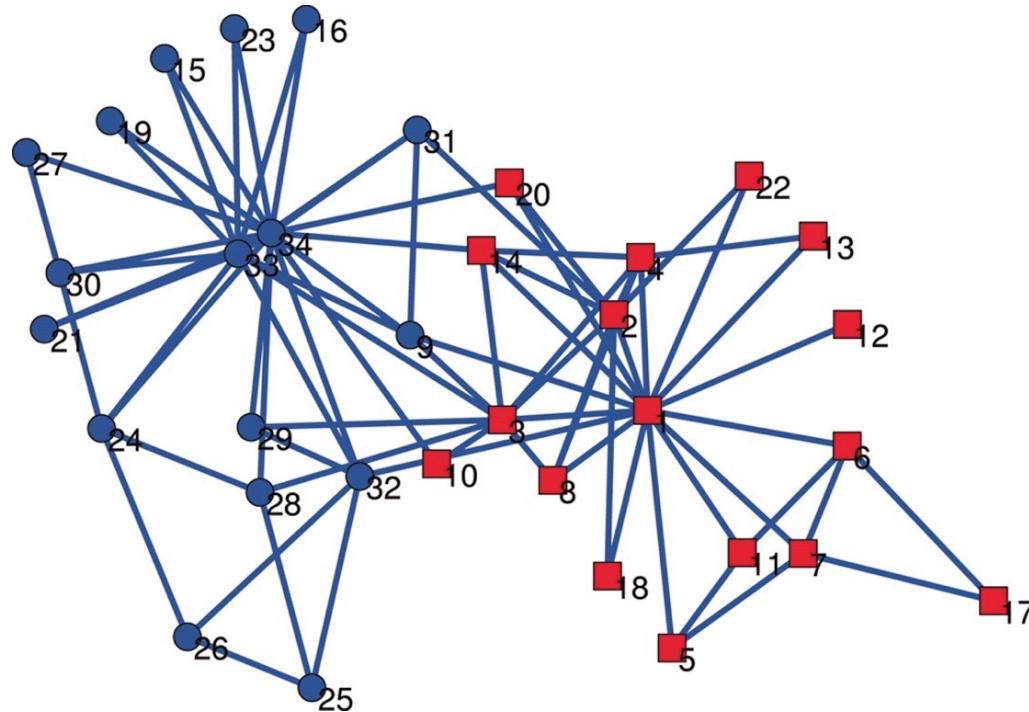
- **Nodos:** color, tamaño, forma, leyenda
- **Enlaces:** color, grosor, curvatura
- **Otros:** comunidades, distribución de la estructura (layout), leyendas, filtros

Colores: HIV risk



Source: Moody J. slides from Social Networks and Health Workshop 2016

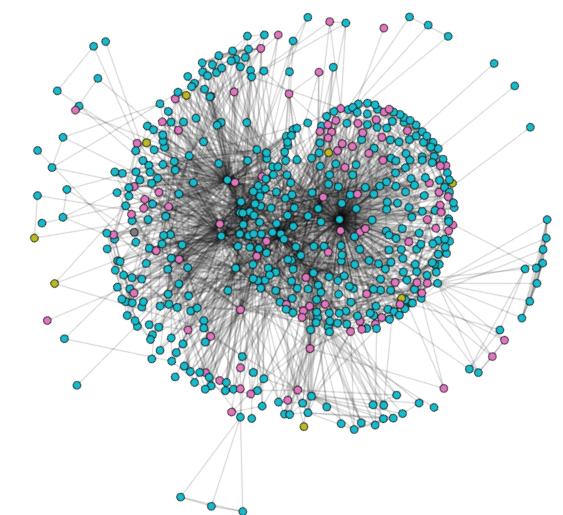
Comunidades: Zachary Karate Club



Zachary, W. W. (1977). *Journal of anthropological research*, 452-473.

Layout: Colaboraciones por género

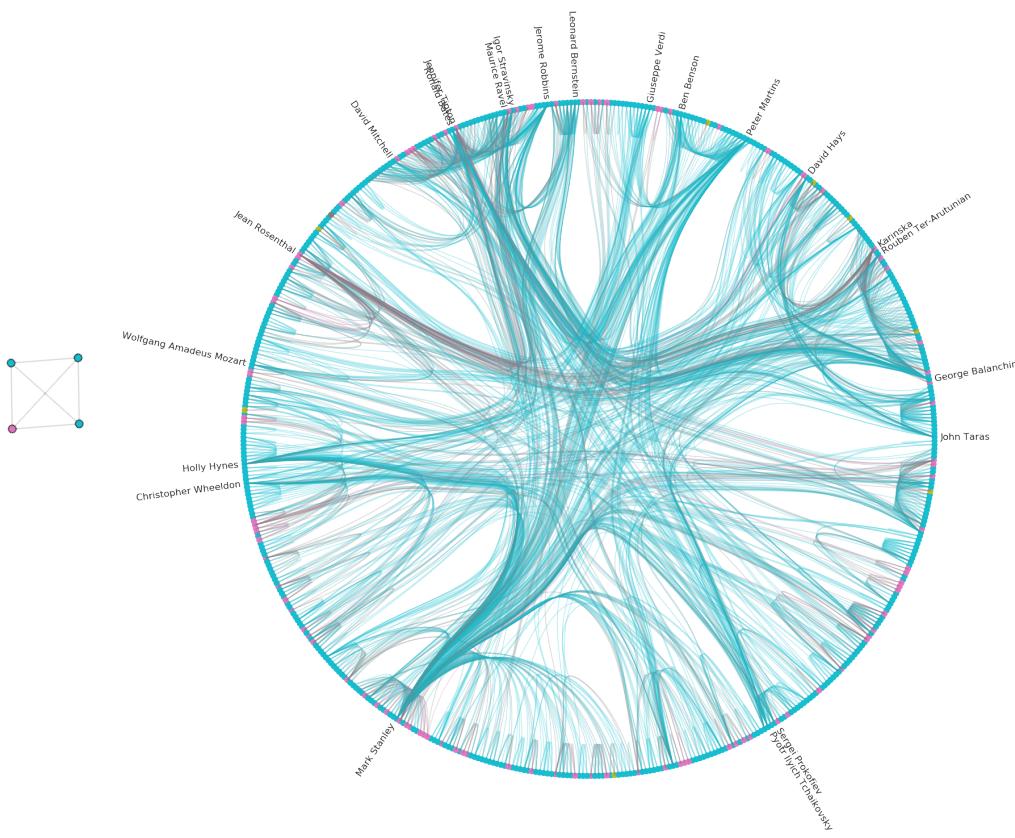
Spring



gender

- female
- male
- unknown
- various

Hierarchical Edge Bundling



Source: Yessica Herrera

- Comunidades, importancia de nodos y tipos de enlaces para monitorear patrones de movilidad urbana

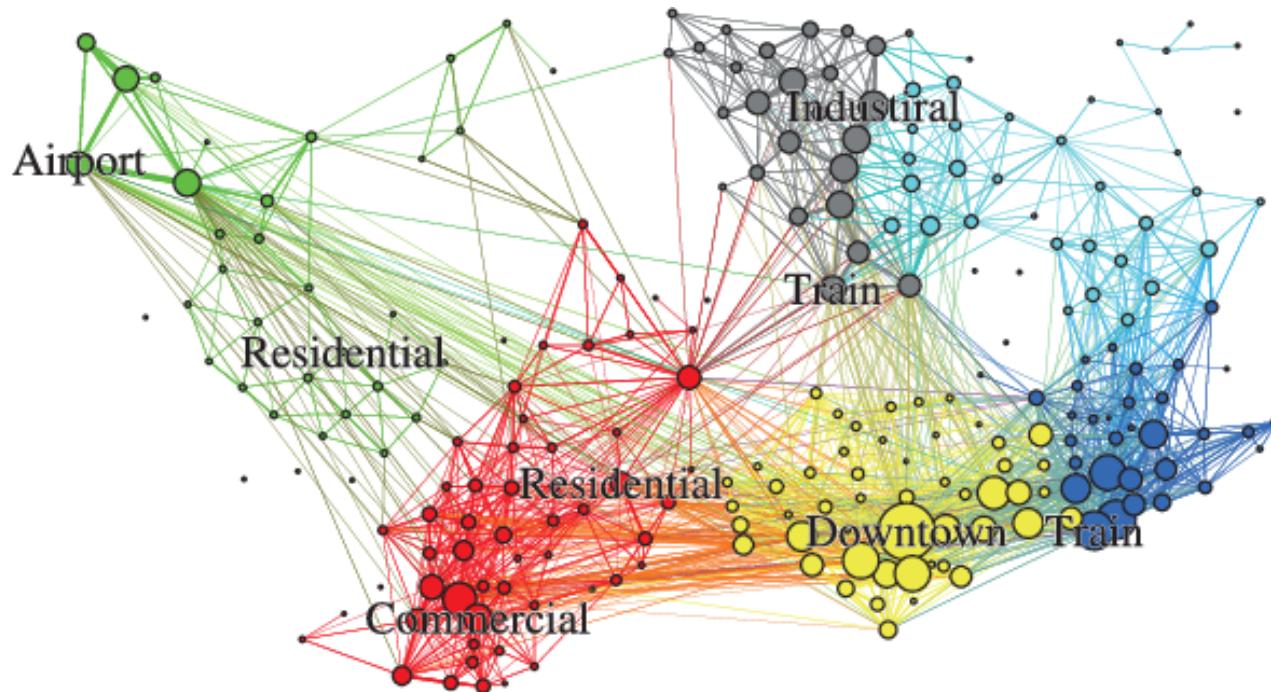


Fig. 2. Human mobility from cellphone data.

Source: Zhang et al. 2018

Preguntas básicas

- Cuáles son las aplicaciones prácticas de redes multinivel y en qué difieren de las redes unipartita?
- Cuál es la utilidad de usar las proyecciones de una red?
- Cuáles son las diferencias estructurales (visibles) en redes de distinta distribución de grado (aleatorias, mundo pequeño, escala libre?)
- Que información *no* se puede obtener de una visualización de red?

Resumen y Conceptos Importantes

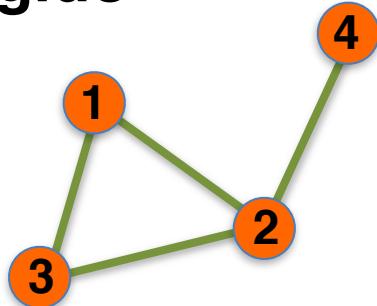
TRES PRINCIPALES CANTIDADES EN NETWORK SCIENCE

Distribución de grado: $P(k)$

Longitud de camino: $\langle d \rangle$

Coeficiente de clustering: $C_i = \frac{2e_i}{k_i(k_i - 1)}$

No-dirigido



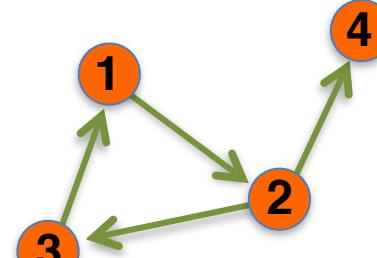
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0$$

$$L = \frac{1}{2} \sum_{i,j=1}^N A_{ij} \quad \langle k \rangle = \frac{2L}{N}$$

Red de actores, interacciones proteína-proteína

Dirigido



$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

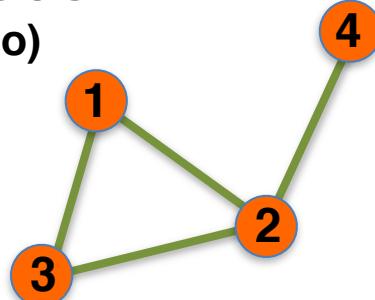
$$A_{ii} = 0$$

$$L = \sum_{i,j=1}^N A_{ij} \quad \langle k \rangle = \frac{L}{N}$$

WWW, red de citas

In=columnas
Out=filas

Sin pesos (no-dirigido)



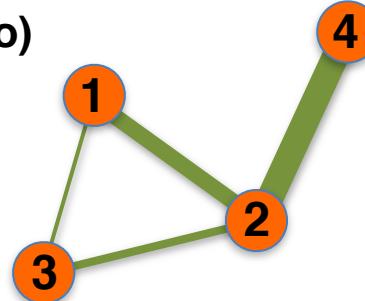
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0$$

$$L = \frac{1}{2} \sum_{i,j=1}^N A_{ij} \quad \langle k \rangle = \frac{2L}{N}$$

Interacciones proteína-proteína, www

Con pesos (no-dirigido)



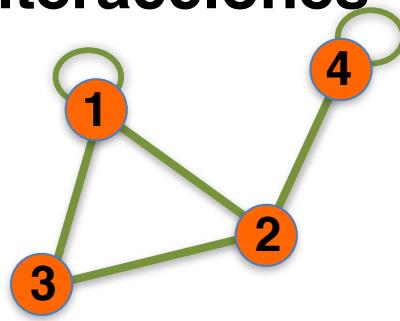
$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0$$

$$L = \frac{1}{2} \sum_{i,j=1}^N \text{nonzero}(A_{ij}) \quad \langle k \rangle = \frac{2L}{N}$$

Grafo de llamadas, red metabólica

Auto-interacciones



$$A_{ij} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

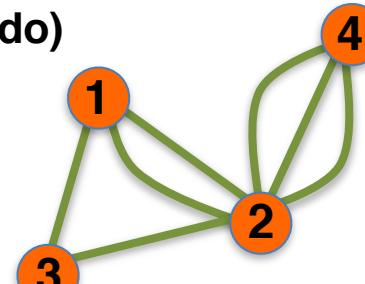
$$A_{ii} \neq 0$$

$$L = \frac{1}{2} \sum_{i,j=1, i \neq j}^N A_{ij} + \sum_{i=1}^N A_{ii}$$



$$A_{ij} = A_{ji}$$

Multigrafo (no-dirigido)



$$A_{ij} = \begin{pmatrix} 0 & 2 & 1 & 0 \\ 2 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix}$$

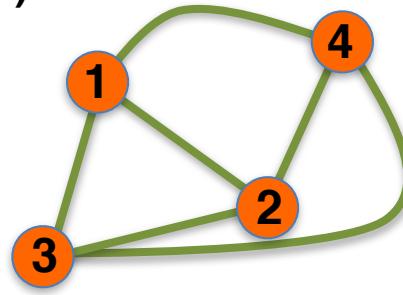
$$A_{ii} = 0$$

$$L = \frac{1}{2} \sum_{i,j=1}^N \text{nonzero}(A_{ij})$$

$$A_{ij} = A_{ji}$$

$$\langle k \rangle = \frac{2L}{N}$$

Grafo completo (no-dirigido)

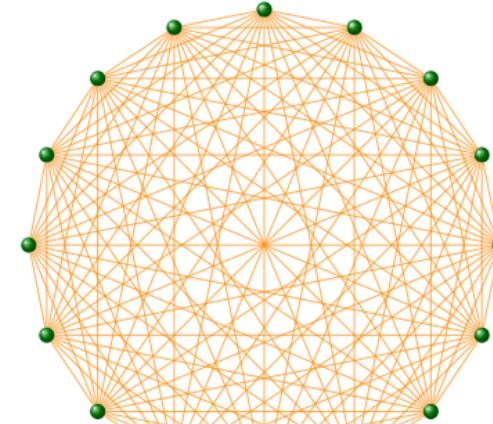


$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \quad A_{i \neq j} = 1$$

$$L = L_{\max} = \frac{N(N-1)}{2} \quad \langle k \rangle = N - 1$$

Red de actores, Interacciones proteína-proteína



GRAPHOLOGY: Las redes reales pueden tener multiples características

WWW > Multigrafo dirigido con auto-interacciones

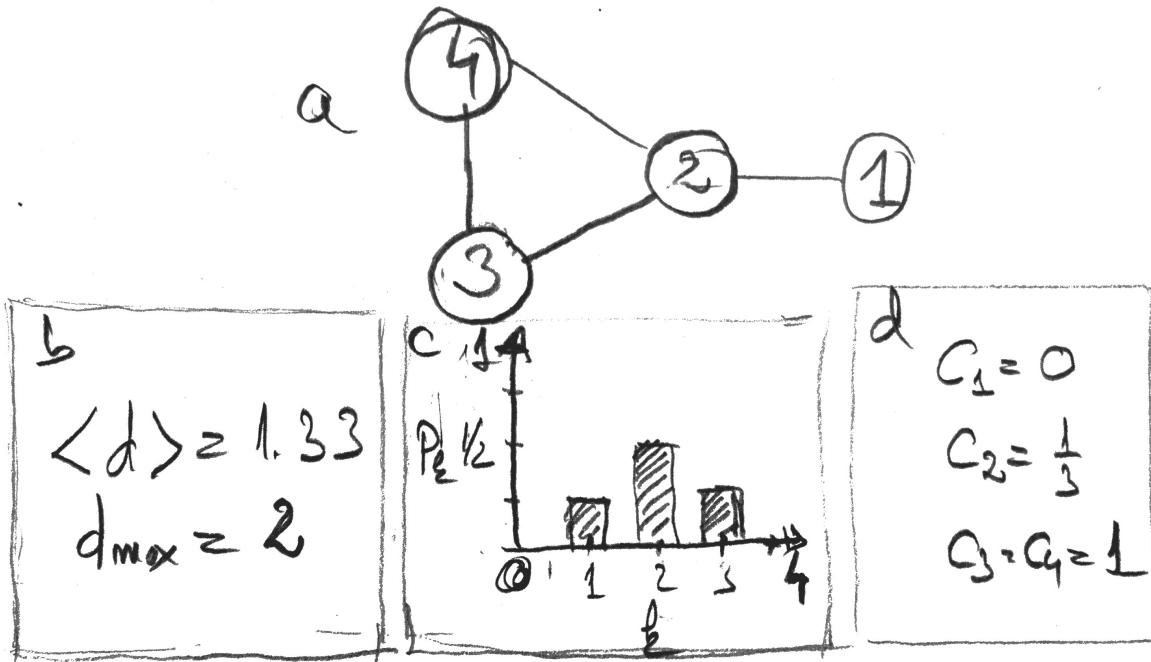
Protein Interactions > no-dirigido, sin pesos, con auto-interacciones

Collaboration network > no-dirigido, multigrafo or con pesos

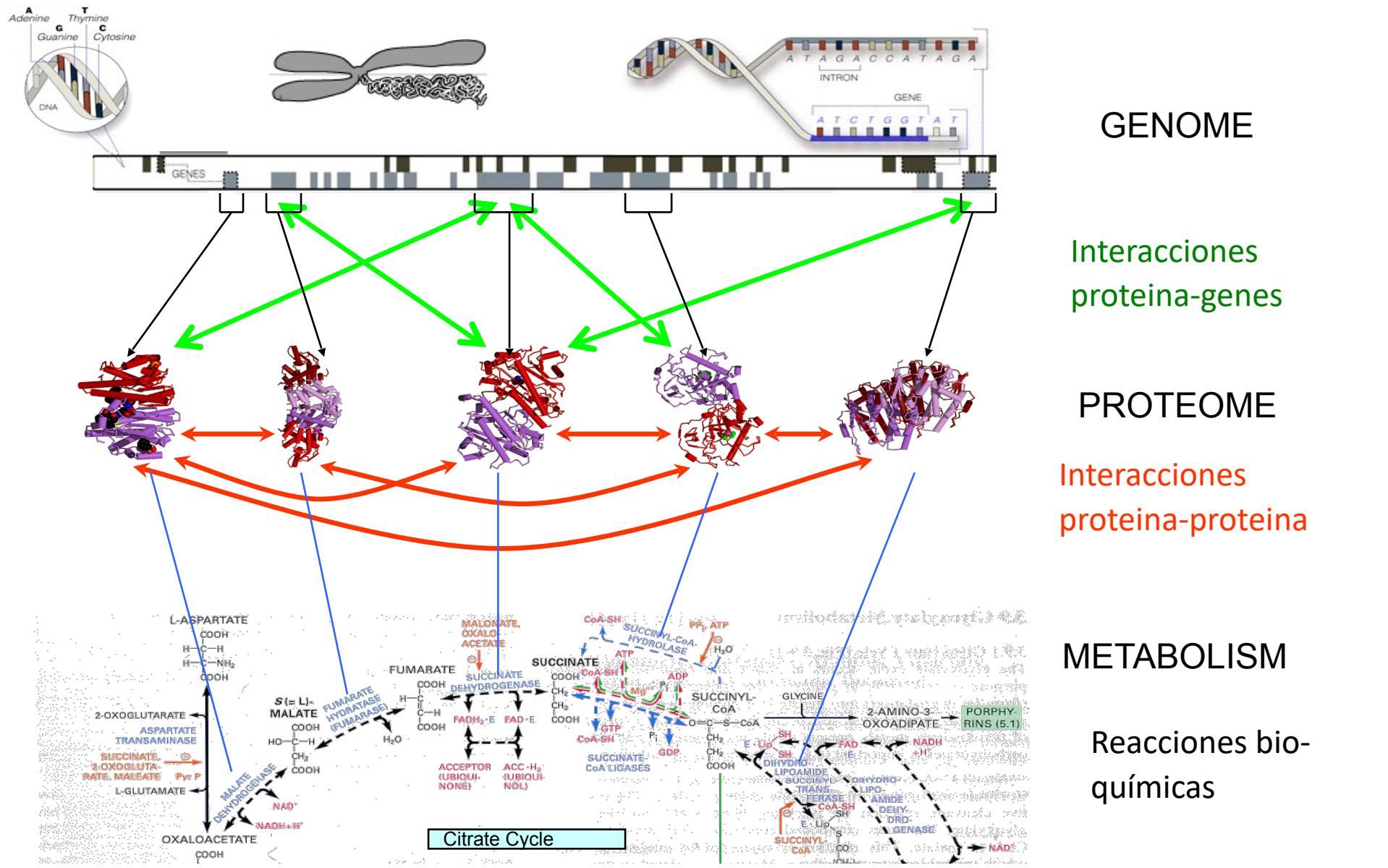
Mobile phone calls > dirigido, con pesos.

Facebook Friendship links > no-dirigido, sin pesos.

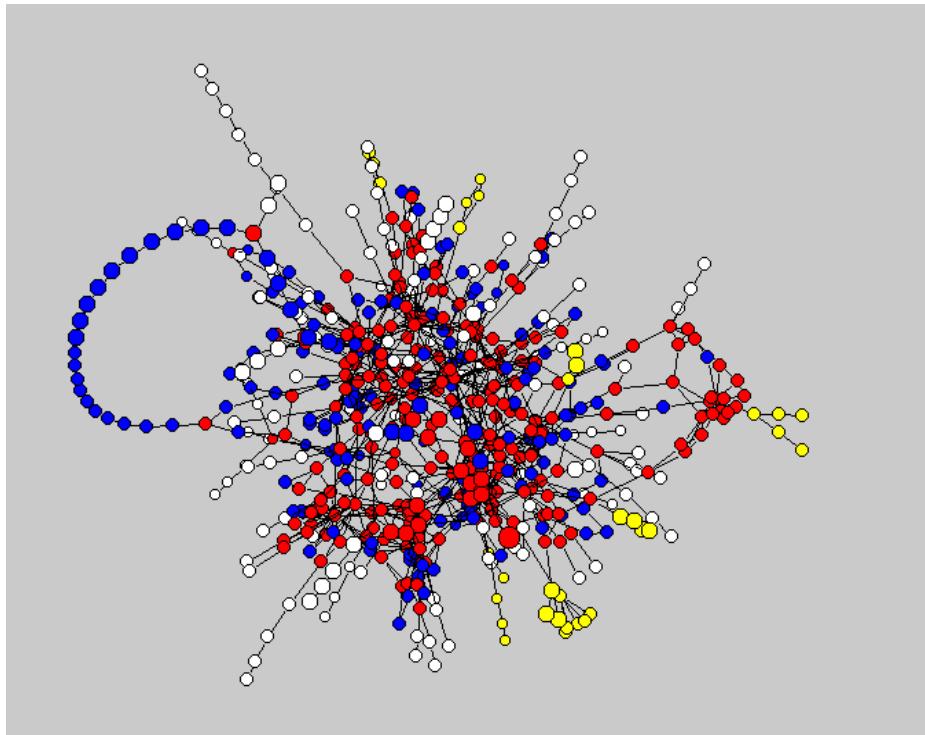
TRES CANTIDADES CENTRALES EN NETWORK SCIENCE



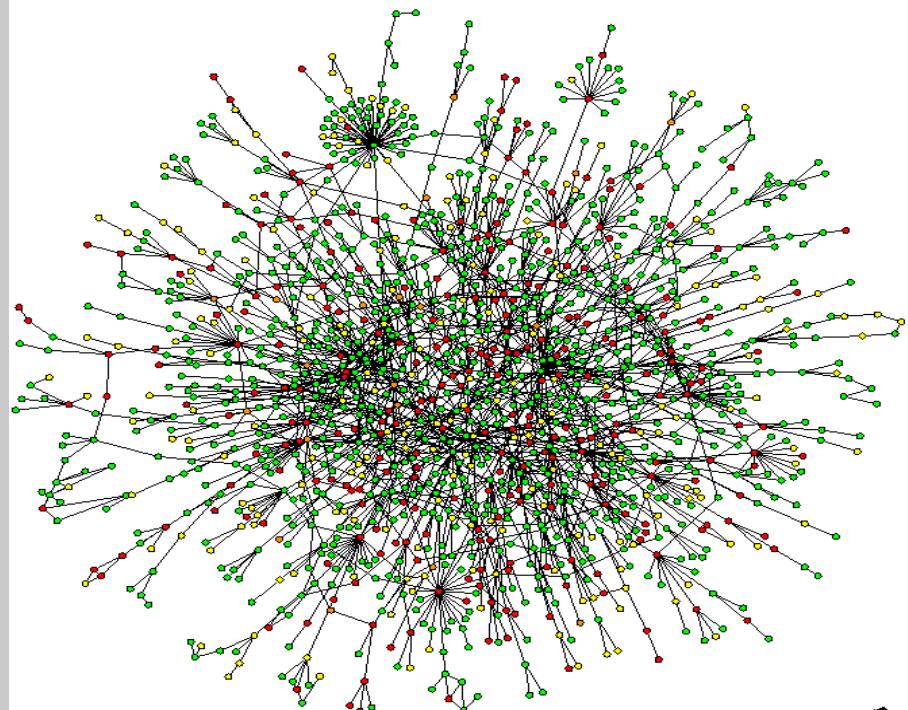
- A. Distribución de grado:** p_k
- B. Longitud de camino:** $\langle d \rangle$
- C. Coeficiente de clustering:** $C_i = \frac{2e_i}{k_i(k_i - 1)}$



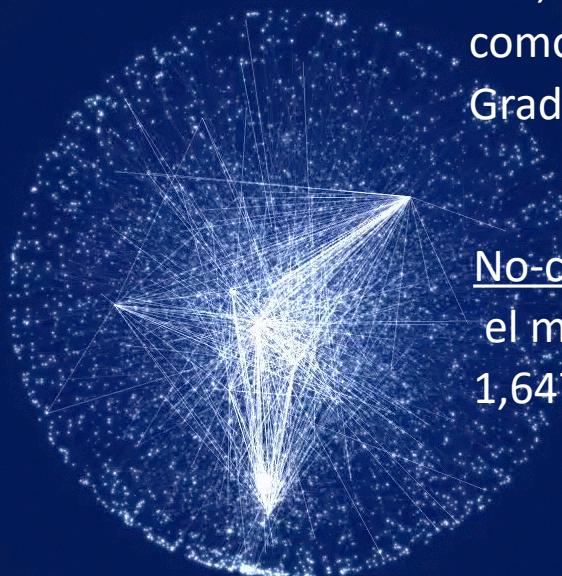
Red metabólica



Interacciones entre Proteinas



A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK



Red no-dirigida

N=2,018 proteinas como nodos

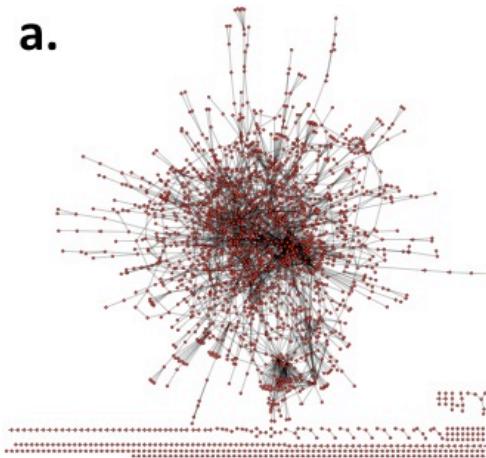
L=2,930 interacciones vinculantes
como links.

Grado promedio $\langle k \rangle = 2.90$.

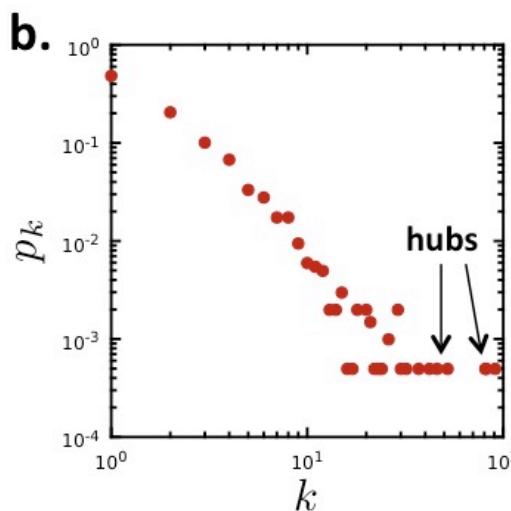
No-conectado: 185 componentes
el más grande (componente gigante)
1,647/2018 nodos

UN CASO DE ESTUDIO: INTERACCION PROTEINA-PROTEINA

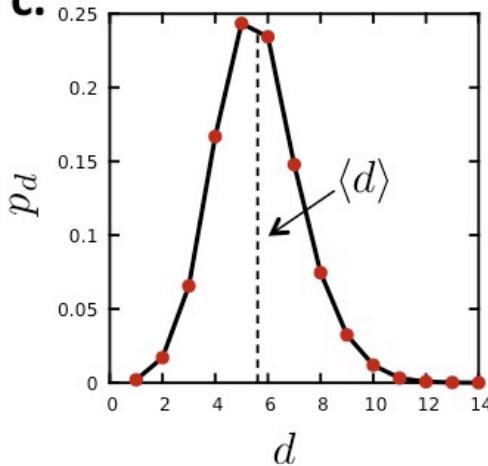
a.



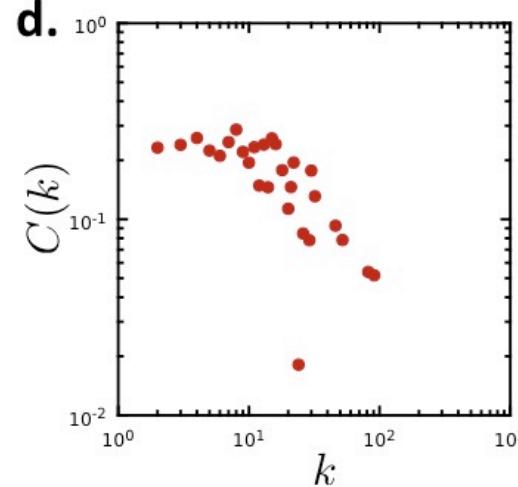
b.



c.



d.



Red no-dirigida

N=2,018 proteínas como nodos

L=2,930 interacciones vinculantes como links.

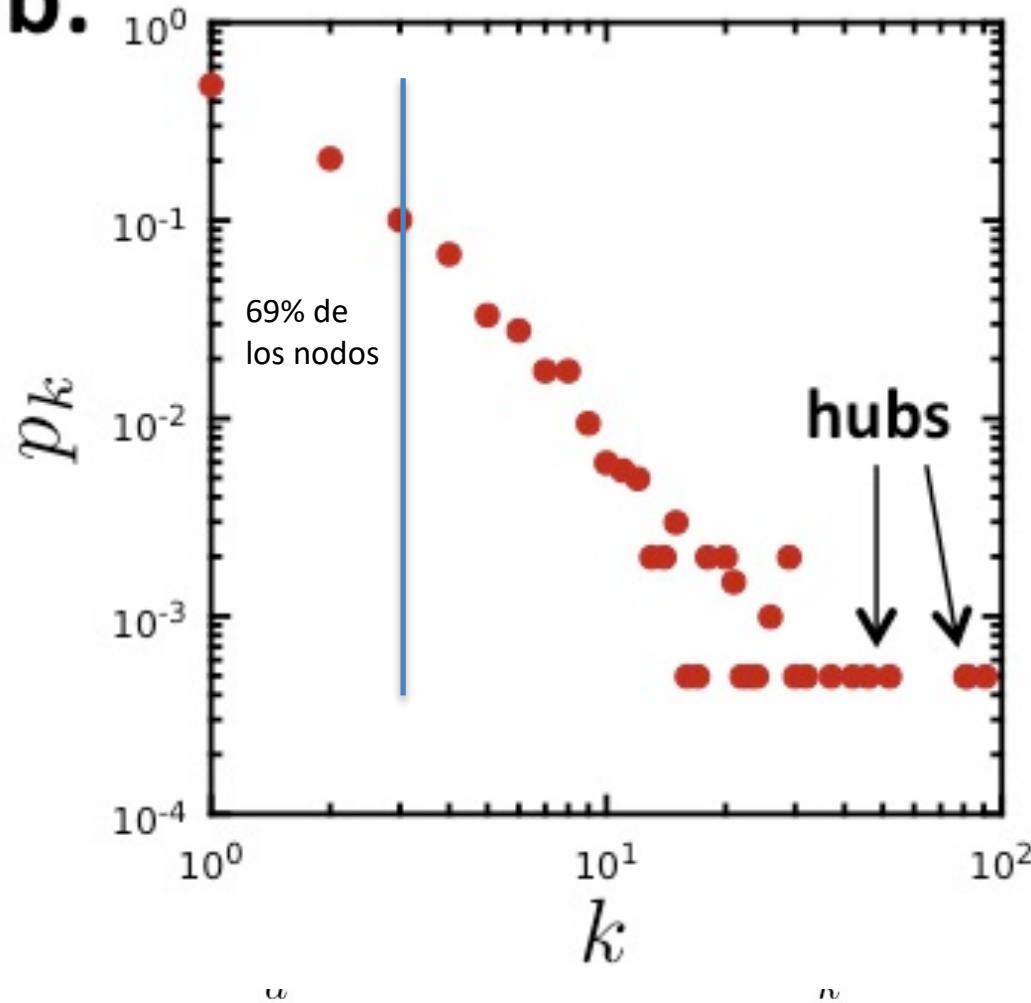
Grado promedio $\langle k \rangle = 2.90$.

No conectado: 185 componentes

las más grande (componente gigante)
1,647 nodos

UN CASO DE ESTUDIO: INTERACCION PROTEINA-PROTEINA

b.



p_k es la probabilidad de que un nodo tenga grado k .

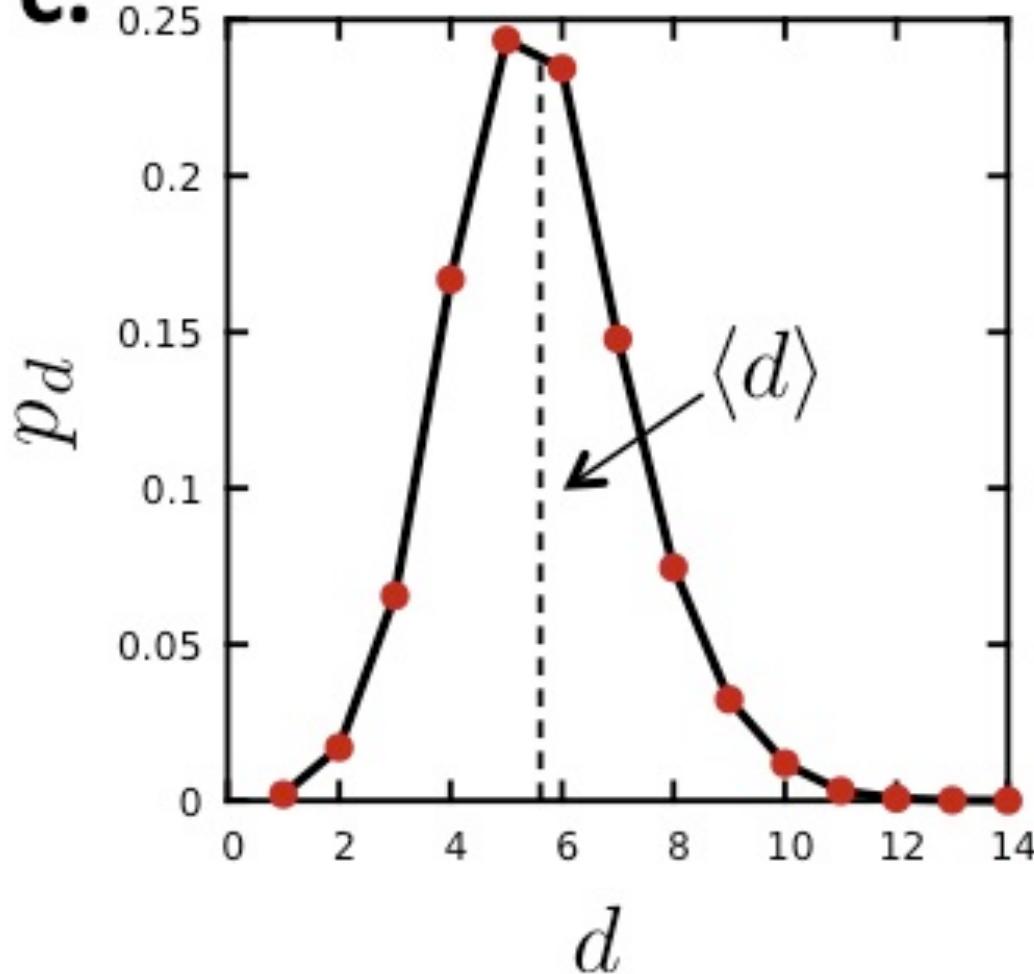
$N_k = \# \text{ nodos con grado } k$

$$p_k = N_k / N$$

Propiedad libre de escala (scale free)

UN CASO DE ESTUDIO: INTERACCION PROTEINA-PROTEINA

C.



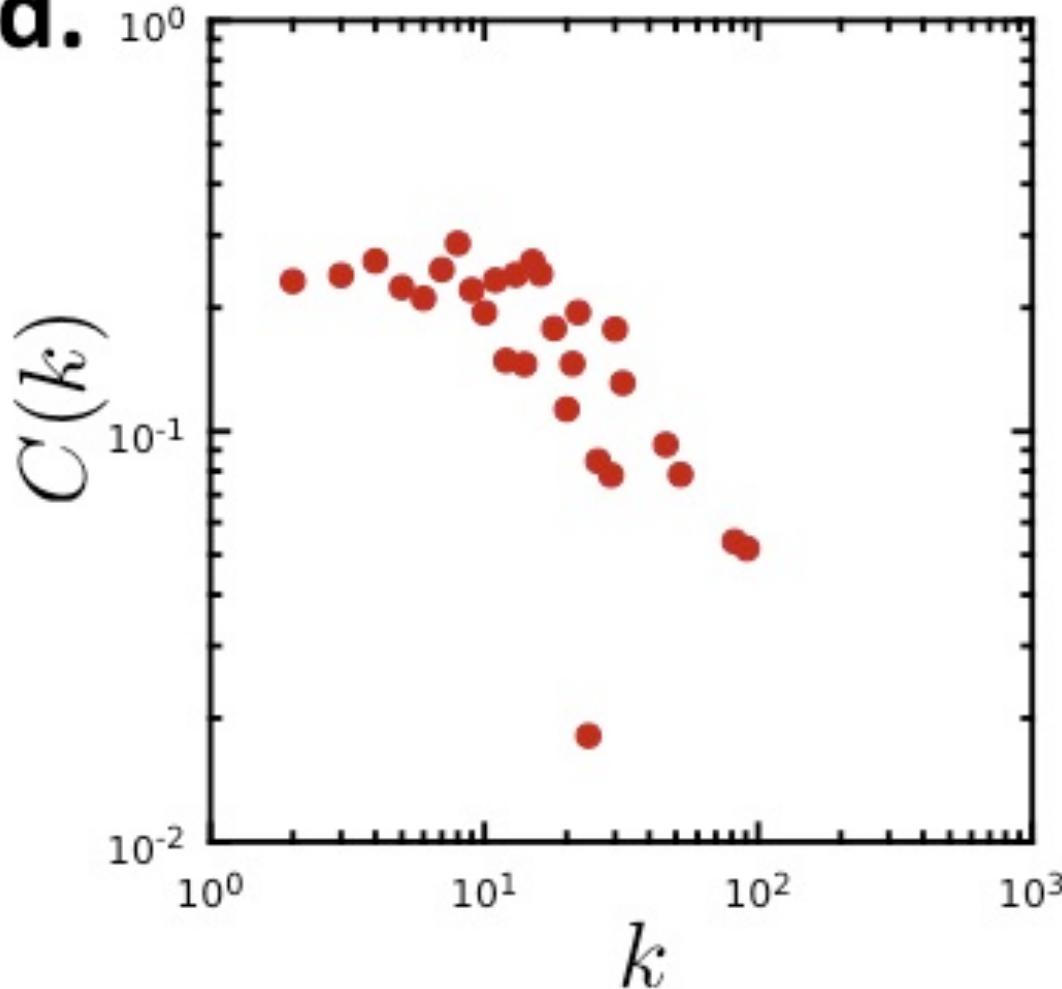
$$d_{\max} = 14$$

$$\langle d \rangle = 5.61$$

Propiedad de mundo pequeño

UN CASO DE ESTUDIO: INTERACCION PROTEINA-PROTEINA

d.



GUIA PARA PROYECTOS FINALES

COMPONENTES DEL PROYECTO

1. ADQUISICIÓN DE DATOS

Descargar la data y ponerla en un formato usable (Ideal, pero en este caso también se proveerán algunos datos de redes en canvas)

2. INFERENCIA Y REPRESENTACIÓN DE LA RED

Qué representan los nodos y los links?

3. Análisis de redes

Qué preguntas quieres responder con esta red, y que herramientas/medidas usarás?

ADQUISICIÓN DE DATOS

- Muchas fuentes de datos en línea tendrán una API (interfaz de programación de aplicaciones) que permite consultar y descargar los datos de forma específica
 - Ejemplo: ¿Cuáles son todas las películas de 1984-1995 protagonizadas por Kevin Bacon y distribuidas por Paramount Pictures?
 - Esto se hace a través de una interfaz web o de una biblioteca dentro de un lenguaje de programación
- Otras fuentes proporcionarán datos en bruto sin procesar (por ejemplo, hojas de cálculo de Excel) que requieren procesamiento, ya sea manualmente o mediante un programa

“GRAPH” ≠ “NETWORK”

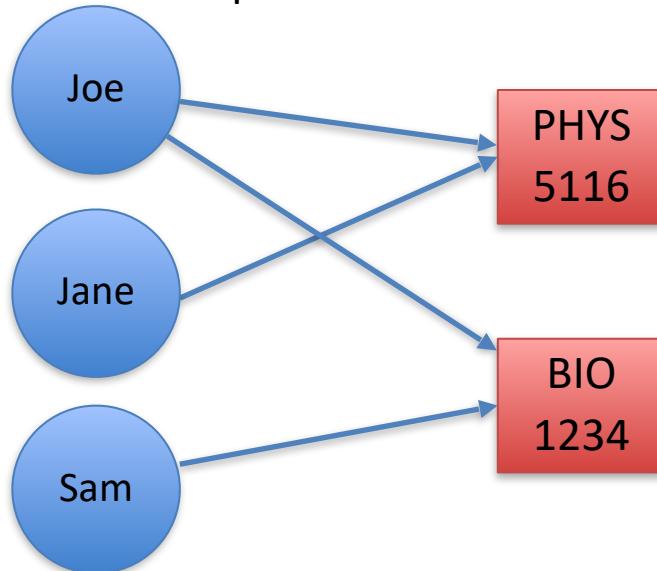
- La mayoría de los conjuntos de datos admitirán más de una representación como una red
- Algunas representaciones serán más o menos informativas que otras.
- ¡Descubrir la "red" que está oculta en sus datos es parte de su proyecto!

RECONSTRUCCION DE RED

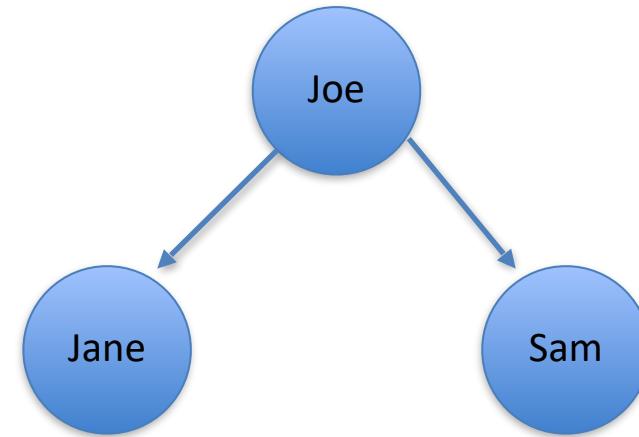
“GRAPH” ≠ “NETWORK”

Supongamos que tiene una lista de estudiantes y los cursos en los que están registrados

Una posible red



Otra posibilidad



Ejemplos

goodreads

Meet your next
favorite book.



- Como IMDB para libros (contiene libros, calificaciones, reseñas, recomendaciones, etc.)
- API disponible en
<https://www.goodreads.com/api>
- Áreas potenciales de investigación:
 - Red de similitud de libros.
 - Detección de comunidades (descubrir géneros)

Global comics database

<http://www.comics.org/>



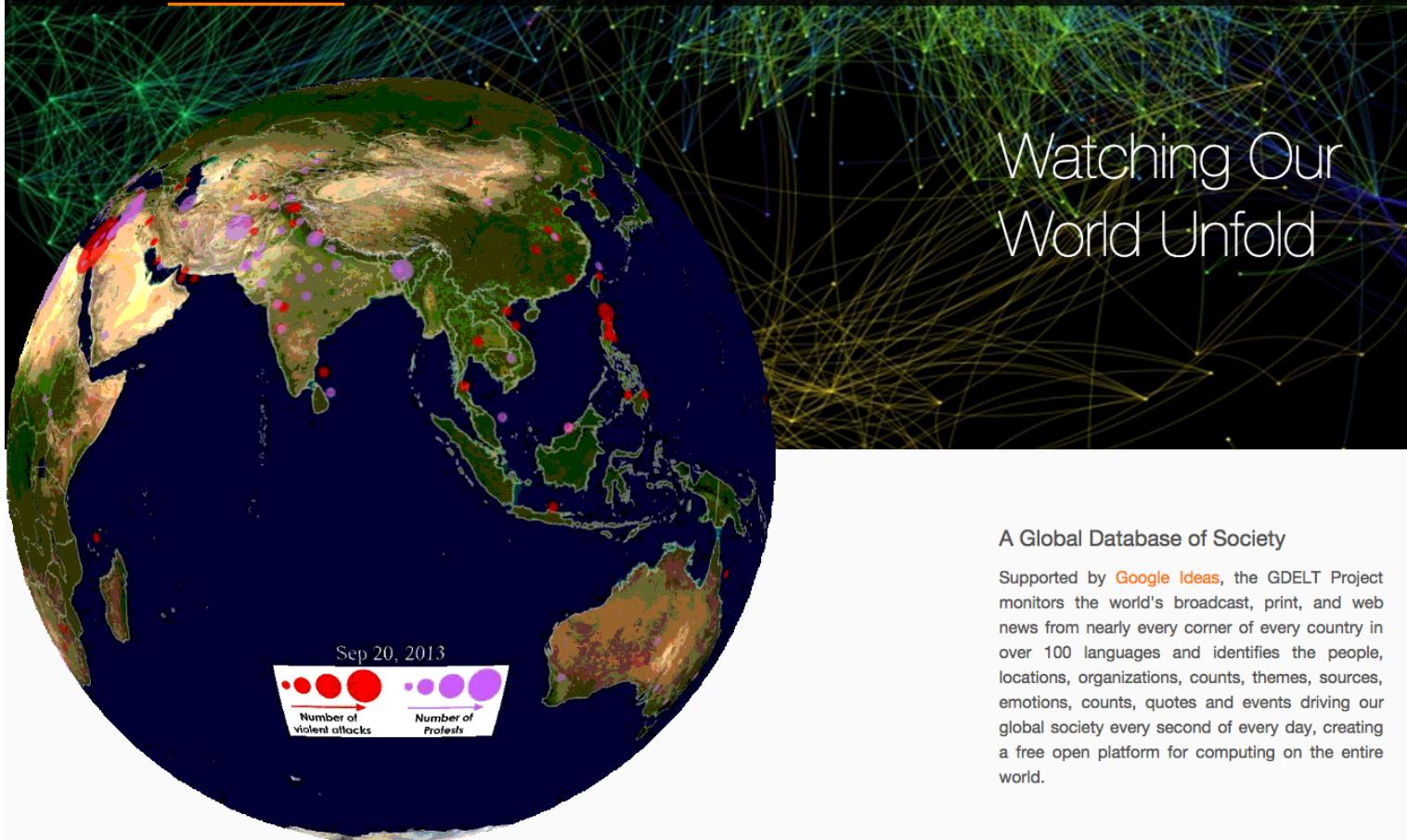
- Muchos datos diferentes sobre cada cómic, por ejemplo:
 - Editor
 - Quien escribió guión / lápiz / tinta
 - Fecha de publicación
- Wiki y la interfaz de búsqueda avanzada disponibles
- Áreas potenciales de investigación:
 - Comics vinculados por personajes comunes.
 - Red de colaboración entre artistas.



<http://www.mendeley.com/>

- Gran base de datos de publicaciones científicas / red social para investigadores.
- API disponible (dev.mendeley.com)
- Idea: utilizar lectores para asignar crédito en co-autoría
 - Los datos consisten en perfiles de usuario + documentos que el usuario ha leído.
 - Las publicaciones (nodos) están vinculadas si ambas están presentes en una o más listas de usuarios
 - Use técnicas desarrolladas recientemente para inferir el crédito de autoría basado en la percepción del usuario:
[\(http://www.pnas.org/content/111/34/12325.abstract\)](http://www.pnas.org/content/111/34/12325.abstract)

The GDELT Project

[Blog](#)[Data](#)[Solutions](#)[About](#)[Intro](#)[Watching](#)[Computing](#)[Downloading](#)[Blogging](#)[Starting](#)

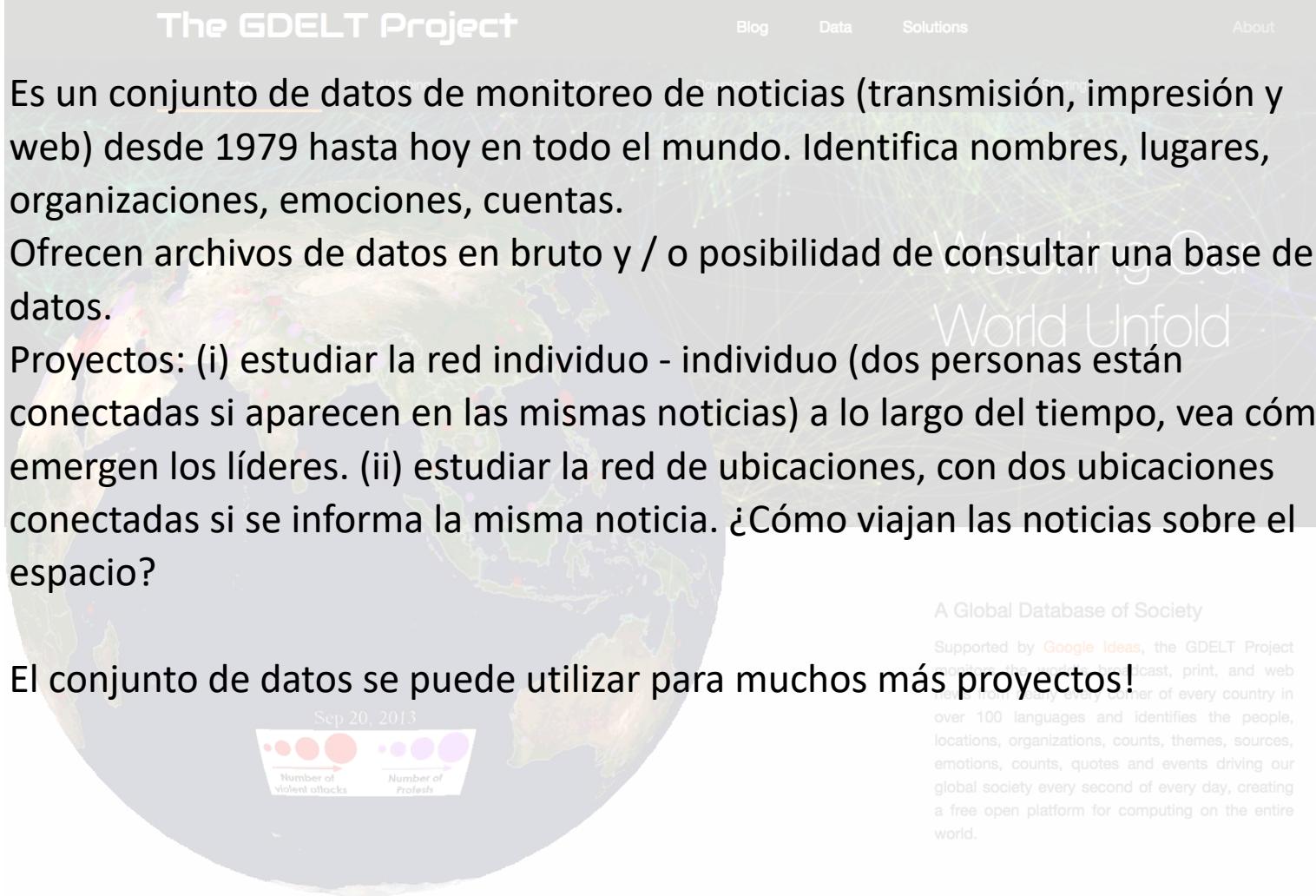
A Global Database of Society

Supported by [Google Ideas](#), the GDELT Project monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organizations, counts, themes, sources, emotions, counts, quotes and events driving our global society every second of every day, creating a free open platform for computing on the entire world.

The GDELT Project

[Blog](#)[Data](#)[Solutions](#)[About](#)

- Es un conjunto de datos de monitoreo de noticias (transmisión, impresión y web) desde 1979 hasta hoy en todo el mundo. Identifica nombres, lugares, organizaciones, emociones, cuentas.
- Ofrecen archivos de datos en bruto y / o posibilidad de consultar una base de datos.
- Proyectos: (i) estudiar la red individuo - individuo (dos personas están conectadas si aparecen en las mismas noticias) a lo largo del tiempo, vea cómo emergen los líderes. (ii) estudiar la red de ubicaciones, con dos ubicaciones conectadas si se informa la misma noticia. ¿Cómo viajan las noticias sobre el espacio?
- El conjunto de datos se puede utilizar para muchos más proyectos!



World Unfold

A Global Database of Society

Supported by Google Ideas, the GDELT Project monitors news media (television, broadcast, print, and web news) in 100 languages, in every corner of every country in over 100 languages and identifies the people, locations, organizations, counts, themes, sources, emotions, counts, quotes and events driving our global society every second of every day, creating a free open platform for computing on the entire world.



Baseball



<http://seanlahman.com/baseball-archive/statistics/>

- Amplia base de datos de estadísticas, a nivel de jugador (estadísticas individuales) y a nivel de equipo (composiciones de equipo, salón de la fama, gerentes, etc.)
- No obstante, posibles direcciones de investigación.
 - ¿Hay características de la red que distinguen a los miembros del Salón de la Fama?
 - Movilidad de jugadores / directivos entre equipos.

Lineamientos finales del proyecto

Medida: N (t), L (t) [t-tiempo si tiene un sistema dependiente del tiempo); P (k) (distribución en grados); $\langle l \rangle$ longitud de camino promedio; C (coeficiente de agrupamiento), C_rand, C (k); Visualización / comunidades; P(w) si tiene una red ponderada; robustez de la red (si corresponde); propagación (si es apropiado).

No es suficiente medir las cosas, es necesario discutir las ideas que ellas ofrecen (significado):

¿Qué aprendiste de cada cantidad que mediste?

¿Cuáles fueron tus expectativas?

¿Cómo se comparan los resultados con tus expectativas?

La restricción de tiempo serán estrictas. Aproximadamente 7min + 3 min preguntas;

No es necesario escribir un informe, basta con entregar la presentación en formato PDF (se recibirán solo PDF).

Deben enviar un email con nombres / título / nombre curso con **24 horas** antes de la presentación.

La primera diapositiva debe contener nombres y título .

Prueba tus diapositivas con el proyector de antemano (es tu responsabilidad comunicar la información de la mayor manera)

Criterio de evaluación:

Uso de herramientas de red (integridad / uso correcto);

Capacidad para extraer información/perspectivas de sus datos utilizando las herramientas de red;

(data != información)

Calidad general del proyecto / presentación.